DOCUMENT RESUME

ED 059 232 TM 001 041

AUTHOR Medley, Donald M.; Norton, Daniel P.
TITLE The Concept of Reliability as it Applies to Behavior Records.
PUB DATE Aug 71
NOTE 12p.; Paper presented at the Annual Meeting of the American Psychological Association, Washington, D. C., August 1971

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Behavior Rating Scales; Behavior Standards; *Classroom Observation Techniques; Classroom Research; Measurement Techniques; *Program Descriptions; *Reliability; *Student Behavior
IDENTIFIERS Personal Record of School Experience

ABSTRACT

A technique for recording pupils' experiences in the classroom, and issues related to the reliability of scores on such records are discussed. The rating instrument, the Personal Record of School Experience (PROSE), is in the form of an optical scanning sheet. Sub-professional personnel can be trained to do the observing, which is recorded on the computer tabulated answer form. (DLG)

ED 059232

THE CONCEPT OF RELIABILITY AS IT APPLIES TO BEHAVIOR RECORDS

Prepared for the 1971 Meetings of the
American Psychological Association in Washington, D.C.

by

Donald M. Medley
University of Virginia

Daniel P. Norton
Educational Testing Service

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

TM 001 170

This paper describes a technique for recording pupils' experiences in the classroom and discusses certain issues related to the reliability of scores on such records. The instrument is called the Personal Record of School Experience or PROSE.

The basic idea behind all versions of OScAR (of which PROSE is the latest) is that important information about classroom processes may be derived from simple frequency counts of readily observable elements of classroom behavior, and that such information can meet the essential requirements for objective measurement. This is primarily achieved by separating the observing and recording of behavior on the one hand from the interpretation and dimensionalization of the records on the other. In other words, the two main steps in the process of behavior measurement are performed by different people at different times.

The recorder's function is to see, discriminate, and record behaviors; it is neither necessary nor desirable for him to have any very clear idea of the significance or meaning of any behavior he records. The only judgments or discriminations he needs to make are those necessary for recognizing which of a set of categories an observed event best fits into. Objectivity is ensured by defining categories so that these discriminations are based (1) on relatively obvious and easily recognized cues, and (2) on cues which are minimally dependent on sophisticated knowledge or on the observer's own set of values. An important byproduct of this is that sub-professional personnel can be trained to do the observations, so that professionally trained personnel are required only for training and supervision.

Objectivity in the second step of the process of behavior measurement--interpreting the record--is assured by turning it over to

the machines. Since PROSE records can be "read" and scored by machines, the only human intervention between the observer's record and the interpretation is to specify in advance how a record is to be interpreted--in other words, to specify the scoring key.

Exhibit 1 is a copy of the instrument itself. You will note that it takes the form of an optical scanning sheet--that is, one which can be read directly onto computer tape by machine. This form was adopted partly to reduce the magnitude of the clerical task involved in coding observational records for analysis, and partly to increase the objectivity of the measurements ultimately derived from the records.

The PROSE recorder enters the classroom early in the morning--before the pupils, if possible--carrying a loose-leaf binder and wearing a compact cassette player over his shoulder. The binder contains one or more PROSE forms for each pupil he plans to observe, arranged in the random order in which he plans to observe them. The cassette player contains a prerecorded tape which emits signals through an earphone at 25 second intervals.

As soon as the pupils arrive the recorder spots the first child on his list and watches what is happening to him. As soon as he hears a signal from the tape, he records whatever event in the child's life is happening at that time. This continues until five events have been recorded, 25 seconds apart. Then the recorder pauses to record the general classroom conditions prevailing during the five events.

The recorder then locates the next child to be observed and repeats the process, until all the children to be observed have been seen once. Then he again observes the first child for another <u>cycle</u>

of five events, and repeats the process as often as time permits or the project requires.

Each event that is recorded is classified on eleven sets of categories, or words, listed on the "statement" side of the form. Three basic kinds of events are recognized: peer contacts, adult contacts, and others. It will be convenient to illustrate the coding process in terms of an event of the second type--an adult contact.

Suppose, for example, that at the time when a signal is heard, the child being observed is sitting quietly on the floor with the rest of the pupils in the class listening to a story the teacher is telling.

The recorder considers the first word, which contains four options. INIT (initiating) is marked if the pupil is seeking to change the kind of attention he is receiving from an adult. STAR is marked if an adult is paying a different kind of attention to the child being observed than to any other child. PART is marked if the adult is giving attention to a group of two or more children of whom the child being observed is one. LSWT (listening or watching) is marked if the child is attending to an adult who is not paying attention to the child. The word is omitted when the child is not in contact with any adult.

The kind of discriminations required of a PROSE recorder are well illustrated by this word--they are based on overt cues which demand very little inferring of intent or effects of behavior. Let us return to our example of the child listening to a story.

The recorder will make a mark after the word PART in column one on the first word because the child is part of a group to which the teacher is attending. If the teacher had been talking to this child only, STAR would have been marked on this word. On the second word, the

recorder marks TCHR (teacher) in the first column to indicate that the adult with whom the child is in contact is the teacher-in-charge. On word 3 he would mark SHTL (show or tell) to indicate that the teacher is showing or telling the children something.

Words 4 and 5 would be left blank because they apply only to peer contacts. Word 6 would be marked after VRBL (verbal) to indicate that the contact was verbal. (This tells us that in this case the teacher is telling, not showing). And so on with the remaining words, which record the child's level of attention, race and sex, activity level, the nature of the task he is performing (if any), and any manifest affect.

When the observer has considered and appropriately marked each of the eleven words in column one, he has recorded what we call a statement about the event in progress when he heard the signal. This process takes from ten to fifteen seconds, so that when the next signal is emitted, the recorder can observe the event in process at that moment.

It is the purpose of the timer and the predetermined random order for observing pupils to ensure that the events observed will be a representative sample of those occurring in the classroom.

The context in which each set of five statements was recorded is coded on the back of the form immediately after the fifth event has been recorded. It indicates such things as what kind of group the child was in and his attitude toward it, what the apparent instructional objectives were and what roles the teacher and each other adult present played, what materials were used, and where in the room the child was. The observer also records any of a number of specified

incidents that may have happened during the cycle of five statements at other times than when the timing signal was given.

The output of each cycle of observations, which takes about five minutes to obtain, consists of five eleven-word statements describing five events plus a description of the context in which they were observed.

The number of different interpretible statements that may be composed using only the eleven words provided is estimated to be more than 200,000; when it is considered that the context in which any statement was made may also vary considerably, it becomes apparent that there is considerable scope for uniqueness and detail in the recording of a single event. And yet the task of the recorder is fairly simple: within the capabilities of a para-professional, for example.

The basic approach to scoring PROSE records is to specify a priori a set of statements describing events which would be expected to occur in the lives of children possessing a certain characteristic of interest--aggressive children, or dependent children, or children in a Montessori program, or whatever. Such a specification is in effect an operational definition of a variable.

The computer is then asked to search each record or set of records and determine the proportion of all statements in the record or set that belong to the specified set of statements. That proportion is the obtained score on the variable in question.

I would like to use the rest of the time allotted to me to discuss some issues related to reliability of observations which have come

up in our work with PROSE. We are particularly concerned about common misuses of coefficients of observer agreement.

Reliability of measurement has to do with how far an obtained score is likely to be from the true score it is supposed to measure. What is this true score in the present case? If we imagine that instead of one recorder observing one child at one time, every recorder in a population of recorders observed all possible children at all possible times, then the proportions of specified statements among all the statements recorded would be the true score in question. This true score, then, is the mean or expected proportion in the population of recorders, times, and children.

In a typical observational study, different recorders observe different children at different times, so the data of the study usually contain all the information needed for reliability estimation. Exhibit 2 is an outline of the analysis of variance of a typical set of data collected to test whether children in different groups differ on some variable of interest. The exhibit also shows how various coefficients may easily be estimated from the mean squares of this analysis.

The common practice of doing a separate reliability study before collecting the main data of an investigation is wasteful and should be discontinued. The only purpose such an undertaking can accomplish is to find out whether a team of recorders has been adequately trained or not--and it is not well adapted to that purpose, either.

The strategy usually followed is to send observers into the field in groups to make simultaneous records of whatever behaviors they may observe, and then to compare the records and calculate

some kind of coefficient of agreement. If the coefficient is too low, the presumption is that the observers need more training.

This may or may not be true. When two observers record the same event differently--when they disagree on how a behavior should be coded--it may be because one or both of them does not know the category definitions; or it may be because the behavior itself is ambiguous, that is, contains elements codable in two or more categories.

Pupils and teachers do not organize their behaviors according to our categories; many of the things they do belong partly in one category and partly in another. If we send enough competent observers to observe such a behavior, some will code it one way, some another. The proportion of behaviors coded in either category will be neither zero nor one but somewhere between, which is where it should be. Brainwashing a team of observers to a point where they would all code the same behavior in the same category would lower their accuracy instead of increasing it!

The way to find out whether a team of recorders is competent is to have them all code a set of filmed or videotaped samples of behavior preselected to contain unambiguous examples of the kinds of behavior they are supposed to record. Any disagreements found in such a situation may be taken as evidence of insufficient training, and near-perfect agreement may be taken as evidence of competence in the system.

Indices of observer agreement in coding the same behaviors have very little to do with the reliability of behavior measurements

anyhow. Since the measurements are usually based on composite frequencies over two or more categories, some observer disagreements count as agreements on one variable and as disagreements on another. And in any case, errors due to observer agreement tend to be negligible in comparison to errors from other sources, such as the instability of the behaviors themselves. Indices of observer agreement have a unique function and their use should be restricted to that function. In any case they should not be cited as evidence of reliability.

PERSONAL RECORD OF SCHOOL EXPERIENCE

Be sure each mark is *dark* and *completely fills* the answer space.
Do not make any stray marks on either side of this sheet.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
▬	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

		CYCLE A	CYCLE B	CYCLE C	
1.	INIT	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	INIT
	STAR	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	STAR
	PART	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	PART
	LSWT	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	LSWT
2.	TCHR	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	TCHR
	AA	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	AA
	TAA	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	TAA
	OBS	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	OBS
	OTH	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	OTH
3.	POS	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	POS
	PRM	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	PRM
	SHTL	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	SHTL
	LSQU	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	LSQU
	DO4	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	DO4
	CNTR	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	CNTR
	NEG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	NEG
4.	AGR	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	AGR
	INIT	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	INIT
	COOP	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	COOP
	WTHD	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	WTHD
	RST	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	RST
5.	AGR	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	AGR
	INIT	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	INIT
	COOP	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	COOP
	WTHD	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	WTHD
	RST	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	RST

		CYCLE A	CYCLE B	CYCLE C	
6.	CNTC	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	CNTC
	MTL	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	MTL
	VRB	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	VRB
7.	OSOG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	OSOG
	OSSG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	OSSG
	SSOG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	SSOG
	SSSG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	SSSG
8.	COOP	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	COOP
	DSTR	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	DSTR
	RIS	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	RIS
	WOA	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	WOA
	DSRP	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	DSRP
9.	HIWL	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	HIWL
	MDWL	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	MDWL
	HINL	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	HINL
	MDNL	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	MDNL
	LOW	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	LOW
10.	FANT	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	FANT
	DVG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	DVG
	CVG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	CVG
	WRK	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	WRK
	KIN	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	KIN
11.	POS	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	POS
	NEG	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	NEG

REMARKS

DO NOT WRITE BELOW THIS LINE

10
Educational Testing Service
Princeton, N. J. 08540
B10BP50D

PERSONAL RECORD OF SCHOOL EXPERIENCE

	A	B	C
3. SHOW			
TELL			
LEAD			
DISCI			
FEER			
MNG			
SPVS			
RSNCE			
HSKP			
INDATT			
LSWT			
NCNT			
4. ONE			
TWO			
3-5			
6+			
ALL			

	A	B	C
5. USED NUMBERS			
USED WORDS			
SNG., TKD TO SLF			
HELPED OTHER P			
COMFORTED P			
SHAFF P			
ASKED P FOR HELP			
REC'D HLP, AFF			
REJECTED BY CHP			
ACCIDENT, HURT			
SHOWED FEAR			
CRIED			
LOST TEMPER			
TATTLED			
LED OTHER P			
BOSSED OTHER P			
WRECKED SOMETHING			
REFUSED HELP			
RESISTED AD			
DISOBEYED			
SHOST TO AD			
RID			
BAD EX			
GOOD EX			
CHORE, ERRAND			
SHAFF AD			
WAITED			

	A	B	C
7. CLSHOT			
TIRET			
LSTMP			
LTTW			
PHRSTR			
8. ART (PAINT, CLAY)			
CRAFT			
MUSICAL INSTR			
BOOKS, WRITING			
PUZZLES, QU GAMES			
BLOCKS			
WHEEL TOYS			
TOOLS, WK BENCH			
SAND			
GYM, EXERCISE			
PET			
CLOTHES, JEWELRY			
FOOD, WATER			
DOLLS			
HOMEMAKING			
CLEANUP TOOLS			
AUDIO VISUAL DEV			
SP INSTRC DEV			
SCIENCE EQUIP			
ASSIGNED SEAT			
TEACHER'S DESK			
OPEN AREA			
TOILET			
PLAYGROUND			
OUT OF ROOM			

EXHIBIT 2

ESTIMATING THE RELIABILITY OF SCORES BASED ON OBSERVATIONAL RECORDS

A study is posited in which N individuals divided into G groups are observed by R recorders a total of V times each. Each recorder observes each individual at least once, and observes every individual the same number of times. This number may vary across recorders--one recorder may see all individuals twice; another may see them three times each. No two recorders observe the same individual at the same time.

The analysis of variance below is the one appropriate for testing for differences in behavior of individuals in different groups, and the formulas show how various coefficients may be estimated from the mean squares obtained in the analysis.

ANALYSIS OF VARIANCE		
Source of Variation	Degrees of Freedom	Mean Square
Between Groups	(G - 1)	a
Between Individuals (in the same group)	(N - G)	b
Between Recorders	(R - 1)	c
Between Observations (made by the same recorder)	(V - R)	d
Interaction, Groups by Recorders	(G-1)(R-1)	e
Interaction, Individuals (in the same group) by Recorders	(N-G)(R-1)	f
Interaction, Groups by Observations (made by the same recorder)	(G-1)(V-R)	g
Residual Variation	(N-G)(V-R)	h
Total Variation	NV - 1	

Coefficient Of:	ESTIMATE	
	For Individuals	For Groups
Reliability (Estimates the correlation between the set of scores actually obtained and a similar set obtained by different recorders observing the individuals at different times)	r=(b-f)/b	r=(a-b-e+g)/a
Stability (Estimates the correlation between the set of scores actually obtained and a similar set obtained by the same recorders observing the individuals at different times)	r=(b-h)/b	r=(a-b+f-g)/a
Observer Agreement (Estimates the correlation between the set of scores actually obtained and a similar set obtained by different recorders observing the individuals at the same times)	r=(b-f)/(b-h)	r=(a-b+f-g)/(a-g)